DOCUMENT RESUME

ED 400 286 TM 025 555

AUTHOR Bergstrom, Betty A.; Gershon, Richard
TITLE Comparison of Item Targeting Strategies for Pass/Fail Computer Adaptive Tests.
PUB DATE Apr 92
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; Algorithms; Comparative Analysis; *Computer Assisted Testing; Higher Education; Item Response Theory; Maximum Likelihood Statistics; Medical Technologists; *Pass Fail Grading; *Test Length
IDENTIFIERS *Item Selection; Rasch Model; Stopping Rules; *Target Populations

ABSTRACT

The most useful method of item selection for making pass-fail decisions with a Computerized Adaptive Test (CAT) was studied. Medical technology students (n=86) took a computer adaptive test in which items were targeted to the ability of the examinee. The adaptive algorithm that selected items and estimated person measures used the Rasch model and a version of maximum likelihood estimation. The stopping rule was based on confidence in the pass/fail decision. Results indicate that when test length is sufficient, targeting items at the ability of the examinee and using a confidence level stopping rule results in the most efficient computer adaptive test for making a pass/fail decision. Examinees whose ability is clearly above or below the pass/fail point then take a minimum number of items, but those whose ability is near the pass point take a test of precision comparable to a test of items targeted at the pass/fail point. An appendix contains an examinee map for the test and a map key. (Contains two tables, three figures, and six references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CAT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
BETTY BERGSTROM

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)

ED 400 286

Comparison of Item Targeting Strategies for Pass/Fail Computer Adaptive Tests

Betty A. Bergstrom

American Society of Clinical Pathologists

Richard Gershon

Computer Adaptive Technologies

Paper presented at the annual meeting
American Educational Research Association, San Francisco, California, 1992.

BEST COPY AVAILABLE

TM025555

Comparison of Item Targeting Strategies
for Computer Adaptive Tests

When a computer adaptive test (CAT) is used to make pass/fail decisions, there are two schools of thought about where items should be targeted.

One point of view, expressed in Kingsbury and Houser (1990), is that items should be targeted to the estimated ability of the examinee. When items are targeted to the ability of the examinee, the information gained from each item is maximized. An adaptive test that provides maximum information should provide the clearest indication of the person's position above or below the pass point. When information is maximized, the standard error of measure is minimized and thus a pass/fail decision can be made with fewer items. Targeting items to the estimated ability of the examinee is more efficient for examinees with abilities well above or below the pass point and should be equally efficient for examinees near the pass point.

A second point of view, expressed by Wainer (1990), is that the most efficient method for determining if a particular person's position is above or below the pass point is to present items whose difficulty matches that of the pass point. The test is adaptive in the sense of the stopping rule implemented. The computer algorithm allows the test to continue until a specified standard error of measure or a specified level of confidence in the pass/fail decision is reached. Wainer believes that the testing process is more efficient when test items center on the pass point and that it is easier to construct an item pool around one level of difficulty rather than across a range of difficulties. In practice, when computer adaptive tests are targeted to examinee ability and examinee ability is not known, items at the beginning of a computer adaptive test are often poorly targeted. Targeting items to the

pass point means that examinees whose ability lies near the pass point receive a test closely targeted to their ability even at the beginning of the test. Since these examinees are the most difficult to make a decision about, targeting items at the pass point will always result in the optimal test for them.

Factors influencing the usefulness of one targeting procedure over the other include the distribution of the examinee population, the stopping rule implemented and the length of the test. Common sense dictates that if the distribution of the examinee population is homogeneous with the median near the pass point there is little reason to target items on ability. For minimum competency tests, however, examinee abilities may be skewed toward the upper end of the distribution. In this case, or in the case of a widely dispersed population, it may be more useful to target to ability since more examinees will take shorter tests.

Test length is an important consideration in the choice of a targeting procedure. Short computer adaptive tests, targeted to examinee ability, may result in misclassification if examinee ability is not well estimated at the beginning of the test or if examinees fail to respond according to model expectations early in the test. An example of this is the high anxious examinee who incorrectly answers several items at the beginning of the test.

The potential advantage of one procedure over the other is determined by the precision with which examinees are measured and the accuracy of the pass/fail decisions rendered. The purpose of this paper is to demonstrate that the most useful method of item selection for making pass/fail decisions with a CAT depends on the above mentioned factors.

Method

Medical technology students from across the country took a computer adaptive test in which the items were targeted to the ability of the examinee. Eighty-six students are included in this study. Other studies using other subsets of students and test conditions are reported elsewhere.

Test Specifications

The adaptive algorithm which selected items and estimated person measures used the Rasch model (Rasch, 1960/1980) and the PROX version of maximum likelihood estimation (Wright and Stone, 1979). The stopping rule was based on confidence in the pass/fail decision. The test stopped when the examinee's estimated ability measure was either 1.3 times the error of measure above the pass point (a clear pass--one tailed 90% confidence interval), or 1.3 times the error of measure below the pass point (a clear fail), or when a maximum test length of 100 items was reached. Minimum test length was 50 items and the pass/fail point was set at .15 logits on the scale.

The CAT ADMINISTRATOR (Gershon, 1989) constructed computer adaptive tests following the content specifications of the traditional paper and pencil certification examination (See Table 1). In the first 50 items, blocks of ten items were administered from subtests 1-4 and blocks of 5 items were administered from subtests 5 and 6. After 50 items, blocks of 4 items (subtests 1-4) and blocks of 2 items (subtests 5 and 6) were administered. Subtest order was selected randomly by the computer algorithm. Items were chosen at random from unused items within .10 logits of the targeted item difficulty within the specified content area.

All examinees started the test with an item whose difficulty was near the pass point (between -.5 and .5 logits). Items were targeted so that

examinees had a 50% probability of correct response and the pass/fail decision was based on the final estimated ability measure. Using the PROX version of maximum liklihood estimation, examinee ability cannot be estimated until the examinee answers at least one item correctly and one item incorrectly. The stepsize for selecting the difficulty of the next item presented, before the examinee ability could be estimated, was 1.00 logit.

Comparison of Targeting

Theoretical standard errors of measure were calculated for examinee abilities from -2.00 logits to +2.00 logits at .05 logit intervals for tests of 50 and 100 items (Wright and Stone, 1979). Theoretical tests were targeted to the pass point (.15), and to examinee ability. The theoretical standard errors of measure were compared with observed standard errors of measure from the computer adaptive test which was targeted to the estimated ability of the examinee.

Effect of Test Length on Pass/Fail Decisions

To examine the impact of giving a short CAT, pass/fail classifications made at 20 items were compared with pass/fail classifications made at the end of the actual CAT. Only examinees for whom a clear decision (90% confidence) was made in less than 100 items were included in this analysis (N=65).

Results

Precision of Measurement

Figure 1 shows that for perfectly targeted 50 item fixed length computer adaptive tests, targeting to the ability of the examinee (SEM=.28) produces lower standard errors of measure than targeting to the pass/fail point (SEM ranges from .28 to .45 logits) for all examinees except those whose ability is

BEST COPY AVAILABLE

at the pass/fail point. Examinees at the pass/fail point have the same SEM regardless of which item selection method is used.

In actual testing situations, perfect targeting is not possible and less than perfect targeting will result in an increased SEM. Figure 2 shows that in the actual CAT, at 50 items, targeting to the ability level of the examinee resulted in a slightly larger SEM for examinees near the pass/fail point than if items had been targeted at the pass/fail point. However, for most examinees, especially those whose ability is relatively far from the pass/fail point, the SEM is considerably smaller than would have been attained if the items had been targeted at the pass/fail point. These examinees took a more efficient test when items were targeted to their current estimated ability than they would have if the items had been targeted to the pass point.

Figure 3 shows the SEM at 100 items for examinees whose ability measure is very near the pass/fail point and thus took a maximum length test. The mean SEM for these 21 examinees is .202 with a standard deviation of .003. Since a perfectly targeted test would yield a SEM of .20, for all practical purposes, the increase in SEM due to poor targeting early in the test has "washed out". For these examinees, targeting either to the pass point or to their ability will provide a comparable result.

Accuracy of the Pass/Fail Decision

When items are targeted to the ability of the examinee, short computer adaptive tests may result in misclassification if the examinee does not respond according to model expectations at the beginning of the test. The examinee map (Gershon, 1989) in the Appendix shows an example of an examinee with a poor start. This examinee missed the first two items and, due to the 1.00 logit stepsize, the difficulty of the third item presented is -1.99

logits. His ability estimate at this point is -1.62. If the test had stopped at 20 items, the examinee would have failed. However his test map shows that he gradually recovers from his initial poor start and at 97 items passes the test with 90% confidence in the decision.

In the observed CAT, thirty-nine (39/86) examinees passed or failed the test with 90% or greater confidence in the accuracy of the decision in a minimum test length of 50 items. Twenty-six (26/86) examinees passed or failed the test in 51 to 99 items with 90% confidence in the accuracy in the decision. Twenty-one (21/86) examinees whose measure was near the pass/fail point took the maximum test length of 100 items and a pass/fail decision was made with less than 90% accuracy. Thus for 65 of the examinees a clear pass/fail decision was reached. Table 2 compares the final pass/fail results with pass/fail results had the test been stopped at 20 items for these 65 examinees. A different pass/fail decision would have been made for 7 (7/65 or 11%) of these examinees had the test been as short as 20 items.

Discussion

Targeting on Ability

If test length is sufficient, targeting items at the ability of the examinee and using a confidence level stopping rule results in the most efficient computer adaptive test for making a pass/fail decision. Examinees whose ability is clearly above or below the pass/fail point take a minimum number of items. Examinees whose ability is near the pass point take a test of comparable precision to a test comprised of items targeted at the pass/fail point.

BEST COPY AVAILABLE

When targeting items to ability, one possible procedure to lessen the effects of a poor start is to reduce the stepsize (used until maximum liklihood ability estimates can be calculated) to a small value (.10 to .20 logits). Another possible procedure is to constrain the difficulty of the first 5 to 10 items to a specified range (e.g. ± .10 logits of the pass/fail point or ± .10 logits of the previous item difficulty) rather than just starting the test with the first item at the pass point. This would limit the possible negative effect of early mistargeting for examinees whose final measure is near the pass point.

In this study the procedure for administering subtests in blocks may have contributed to inaccuracy of decision at 20 items. If an examinee's ability was inconsistent across subtests, his performance on the first subtest had a great impact on the pass/fail decision at 20 items. For example, an examinee who did very well on subtest 1, but performed poorly on other areas of the test, would have passed at 20 items but failed the test. A better procedure might be to distribute items across subtests rather than in blocks or to use smaller blocks of items for each subtest.

Targeting on the Pass/Fail Point

If a computer adaptive test is a short test, placing items at the pass/fail point and using a stopping rule that requires a specified level of precision (SEM) may be a useful combination of procedures. While examinees whose ability is far from the pass/fail point may take additional items to reach the specified SEM, examinees near the pass/fail point, for whom the decision is most difficult to make, will be presented with well targeted items even at the start of the test and will thus be measured with the greatest precision.

References

Gershon, R.C. (1989). CAT ADMINISTRATOR (Computer Program) Chicago: Computer Adaptive Technologies.

Gershon, R.C. (1991). CAT MAP PROGRAM (Computer Program) Chicago: Computer Adaptive Technologies.

Kingsbury, G.G. and Houser, R.L. (1990, April). Assessing the utility of item response models: Computerized adaptive testing. Paper presented to the annual meeting of the National Council on Measurement in Education, Boston, MA.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Wainer, H. (1990). Four Potholes on the road to a CAT version. Educational Testing Service, Princeton, NJ.

Wright, B.D. and Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.

BEST COPY AVAILABLE

Appendix

Map Key

A. Summary Statistics

1. Total Test/Subtest
   - T - total
   - M - microbiology
   - BB - blood banking
   - C - chemistry
   - H - hematology
   - BF - body fluids
   - I - immunology

B. Number of items

C. Number of items correct

D. Number of items incorrect

E. Ability measure

F. Error of measure

G. Average item difficulty

H. Sum of the squares (item difficulty)

I. Item number

J. Subtest identifier

K. Item difficulty

L. Response selected

M. Right/wrong 0-incorrect 1-correct

N. Current estimated ability measure

O. Current estimated error

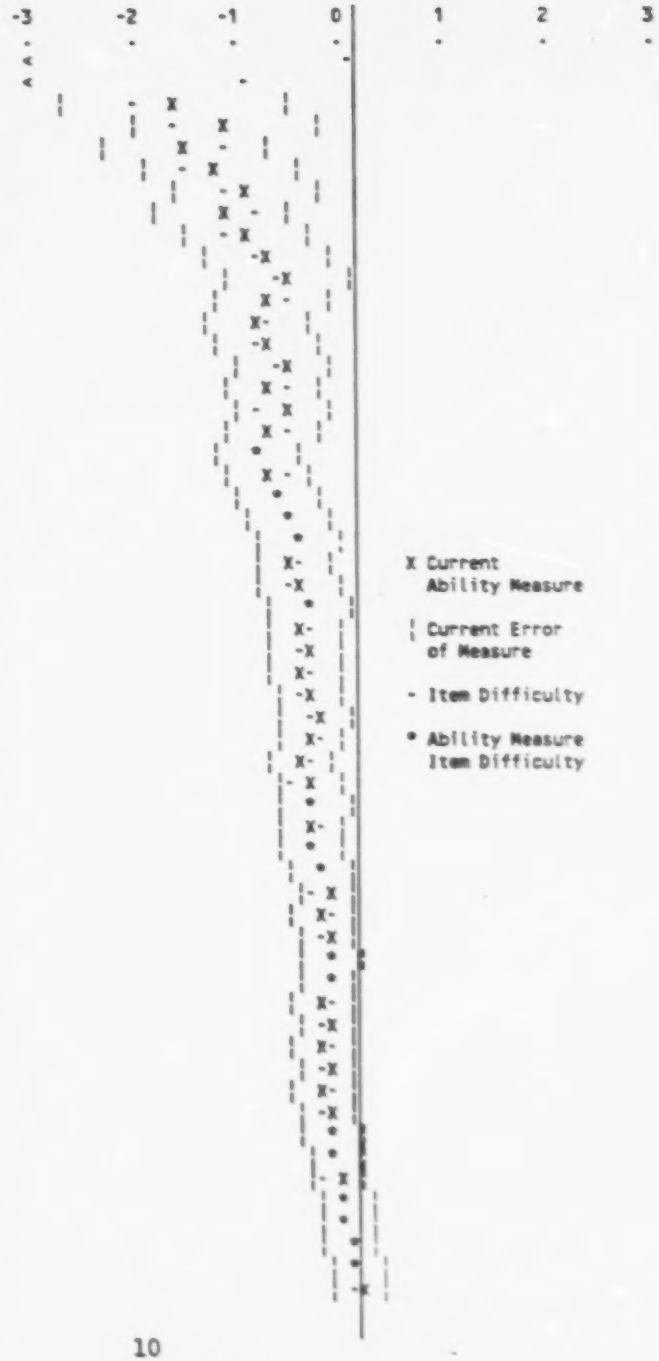
P. Time (in seconds)

Pass/fail point (.15 logits)

Examinee Map

A	B	C	D	E	F	G	H
T	97	62	35	0.43	0.21	-0.14	26.0
M	22	15	7	0.42	0.46	-0.34	16.0
BB	21	15	6	0.91	0.48	-0.00	1.7
C	18	10	8	0.28	0.47	0.06	1.1
W	18	9	9	-0.26	0.47	-0.26	4.8
BF	9	7	2	1.10	0.80	-0.15	0.9
I	9	6	3	0.53	0.71	-0.16	1.5

	I	J	K	L	M	N	O	P
1	81	M	0.13	2	0	-9.99	-9.99	18
2	33	M	-0.93	2	0	-9.99	-9.99	174
3	4	M	-1.99	1	1	-1.62	1.22	4
4	7	M	-1.62	3	1	-1.10	1.00	21
5	20	M	-1.08	2	0	-1.50	0.91	57
6	8	M	-1.55	2	1	-1.17	0.82	31
7	17	M	-1.14	2	1	-0.88	0.76	40
8	34	M	-0.82	1	0	-1.12	0.71	109
9	19	M	-1.13	1	1	-0.90	0.67	25
10	35	M	-0.79	2	1	-0.69	0.65	24
11	489	M	-0.64	2	1	-0.49	0.63	9
12	495	M	-0.68	1	0	-0.67	0.59	67
13	488	M	-0.73	2	0	-0.83	0.56	55
14	485	M	-0.83	1	1	-0.68	0.54	7
15	490	M	-0.59	4	1	-0.54	0.53	97
16	491	M	-0.53	2	0	-0.67	0.50	64
17	487	M	-0.76	4	1	-0.55	0.49	18
18	492	M	-0.53	1	0	-0.67	0.47	29
19	486	M	-0.82	2	0	-0.78	0.46	50
20	493	M	-0.53	1	1	-0.67	0.45	56
21	670	I	-0.59	4	1	-0.57	0.44	50
22	671	I	-0.53	3	1	-0.47	0.43	27
23	675	I	-0.43	2	1	-0.38	0.43	12
24	677	I	-0.37	3	0	-0.47	0.41	44
25	674	I	-0.52	3	1	-0.39	0.41	21
26	614	BF	-0.34	4	1	-0.30	0.40	27
27	615	BF	-0.29	1	0	-0.38	0.39	22
28	613	BF	-0.41	4	1	-0.31	0.39	34
29	616	BF	-0.28	4	0	-0.38	0.38	28
30	612	BF	-0.42	2	1	-0.31	0.37	15
31	199	BB	-0.29	1	1	-0.24	0.37	40
32	202	BB	-0.24	1	0	-0.31	0.36	161
33	198	BB	-0.34	2	0	-0.37	0.35	28
34	195	BB	-0.46	2	1	-0.32	0.35	34
35	200	BB	-0.28	2	1	-0.26	0.35	16
36	203	BB	-0.22	4	0	-0.31	0.34	43
37	197	BB	-0.34	2	1	-0.26	0.33	20
38	204	BB	-0.16	3	1	-0.20	0.33	46
39	201	BB	-0.26	3	1	-0.15	0.33	35
40	207	BB	-0.11	1	0	-0.20	0.32	101
41	361	C	-0.20	4	1	-0.15	0.32	66
42	367	C	-0.11	2	1	-0.10	0.32	33
43	371	C	-0.06	4	0	-0.15	0.31	32
44	366	C	-0.15	4	0	-0.20	0.31	66
45	362	C	-0.20	2	1	-0.15	0.30	36
46	368	C	-0.11	4	0	-0.19	0.30	52
47	360	C	-0.21	2	1	-0.15	0.30	66
48	365	C	-0.15	3	0	-0.19	0.29	85
49	359	C	-0.22	2	1	-0.15	0.29	11
50	369	C	-0.11	1	1	-0.11	0.29	45
51	619	BF	-0.06	2	1	-0.07	0.29	15
52	618	BF	-0.20	2	1	-0.03	0.29	12
53	220	BB	-0.01	1	1	0.01	0.28	28
54	225	BB	0.03	1	1	0.05	0.28	31
55	226	BB	0.08	4	1	0.09	0.28	38
56	229	BB	0.10	4	1	0.13	0.28	7
57	522	M	0.08	4	1	0.16	0.28	9



BEST COPY AVAILABLE

58	526	H	0.17	4	0	0.13	0.27	9
59	521	H	0.08	1	0	0.09	0.27	53
60	523	H	0.13	4	1	0.12	0.27	12
61	690	I	0.10	1	1	0.16	0.27	29
62	692	I	0.22	3	0	0.12	0.26	70
63	80	K	0.07	4	1	0.16	0.26	17
64	84	H	0.17	4	1	0.19	0.26	24
65	87	H	0.22	4	1	0.22	0.26	16
66	88	H	0.24	3	1	0.26	0.26	34
67	392	C	0.25	2	1	0.29	0.26	25
68	393	C	0.30	2	1	0.32	0.26	39
69	398	C	0.36	1	1	0.35	0.26	14
70	397	C	0.31	1	0	0.32	0.25	41
71	632	BF	0.25	4	1	0.35	0.25	9
72	633	BF	0.36	2	1	0.38	0.25	38
73	244	BB	0.36	1	1	0.41	0.25	36
74	246	BB	0.42	4	0	0.38	0.25	30
75	243	BB	0.35	3	0	0.35	0.24	55
76	242	BB	0.31	3	1	0.38	0.24	126
77	396	C	0.30	4	0	0.35	0.24	56
78	399	C	0.40	3	0	0.32	0.24	73
79	395	C	0.30	2	1	0.35	0.24	11
80	400	C	0.41	3	0	0.32	0.23	14
81	94	H	0.31	1	0	0.29	0.23	56
82	91	H	0.27	4	1	0.32	0.23	75
83	95	H	0.35	3	1	0.34	0.23	11
84	93	H	0.31	1	1	0.37	0.23	87
85	695	I	0.36	1	0	0.34	0.23	11
86	694	I	0.31	3	1	0.37	0.22	7
87	539	H	0.36	1	0	0.34	0.22	35
88	537	H	0.31	4	0	0.32	0.22	53
89	538	H	0.35	4	1	0.34	0.22	40
90	536	H	0.30	3	1	0.36	0.22	82
91	97	H	0.36	3	0	0.34	0.22	44
92	92	H	0.30	4	1	0.36	0.22	27
93	98	H	0.37	3	1	0.39	0.22	30
94	99	H	0.40	2	0	0.36	0.21	44
95	241	BB	0.31	3	1	0.38	0.21	57
96	245	BB	0.41	2	1	0.41	0.21	12
97	240	BB	0.31	4	1	0.43	0.21	24

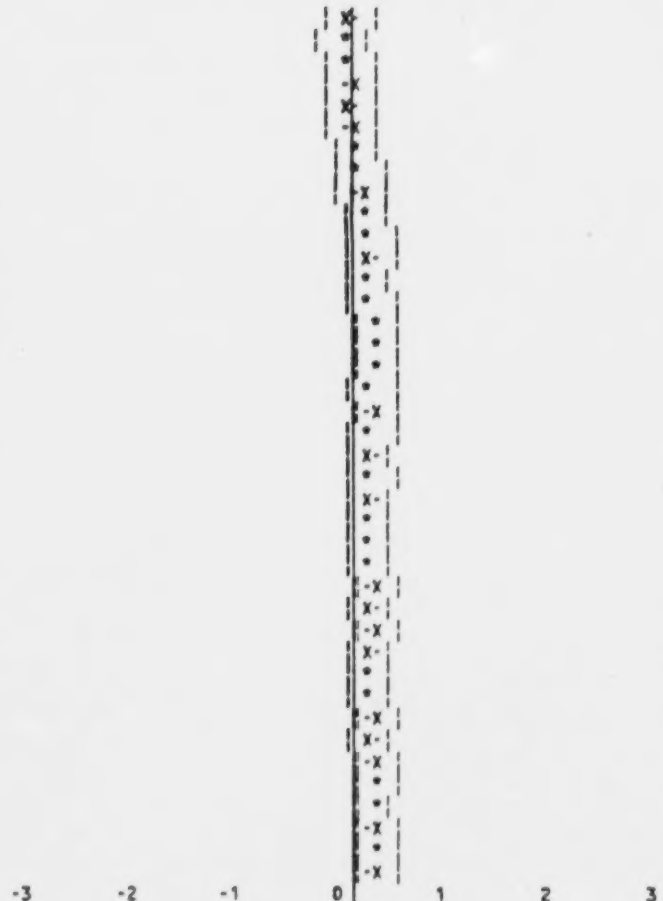


Table 1
Item Bank Description

Subtest	Test Plan Distribution*	Number of Items in Bank	Easiest Item	Mean	Hardest Item	SD
Microbiology	20%	147	-2.89	-.06	2.38	.96
Blood Banking	20%	165	-2.21	-.07	2.94	1.00
Chemistry	20%	142	-3.61	-.07	2.97	1.06
Hematology	20%	135	-2.80	-.05	2.97	.97
Body Fluids	10%	72	-2.24	-.09	3.84	.97
Immunology	10%	65	-2.78	.25	2.04	.96
Bank Scale	100%	726	-3.61	-.02	3.84	1.00

* The test plan distribution for computer adaptive tests was the same as the test plan for the traditional fixed length written certification examination.

Table 2

Pass/Fail Consistency *
Comparison of Decision after 20 Items
and Final Decision

Decision at 20 Items

Final Decision		Pass	Fail
	Pass	38	5
	Fail	2	20

N=65

* For examinees for whom a clear (90% confidence) decision was reached. 11% of the examinees would have been affected by a short test.

Figure 1

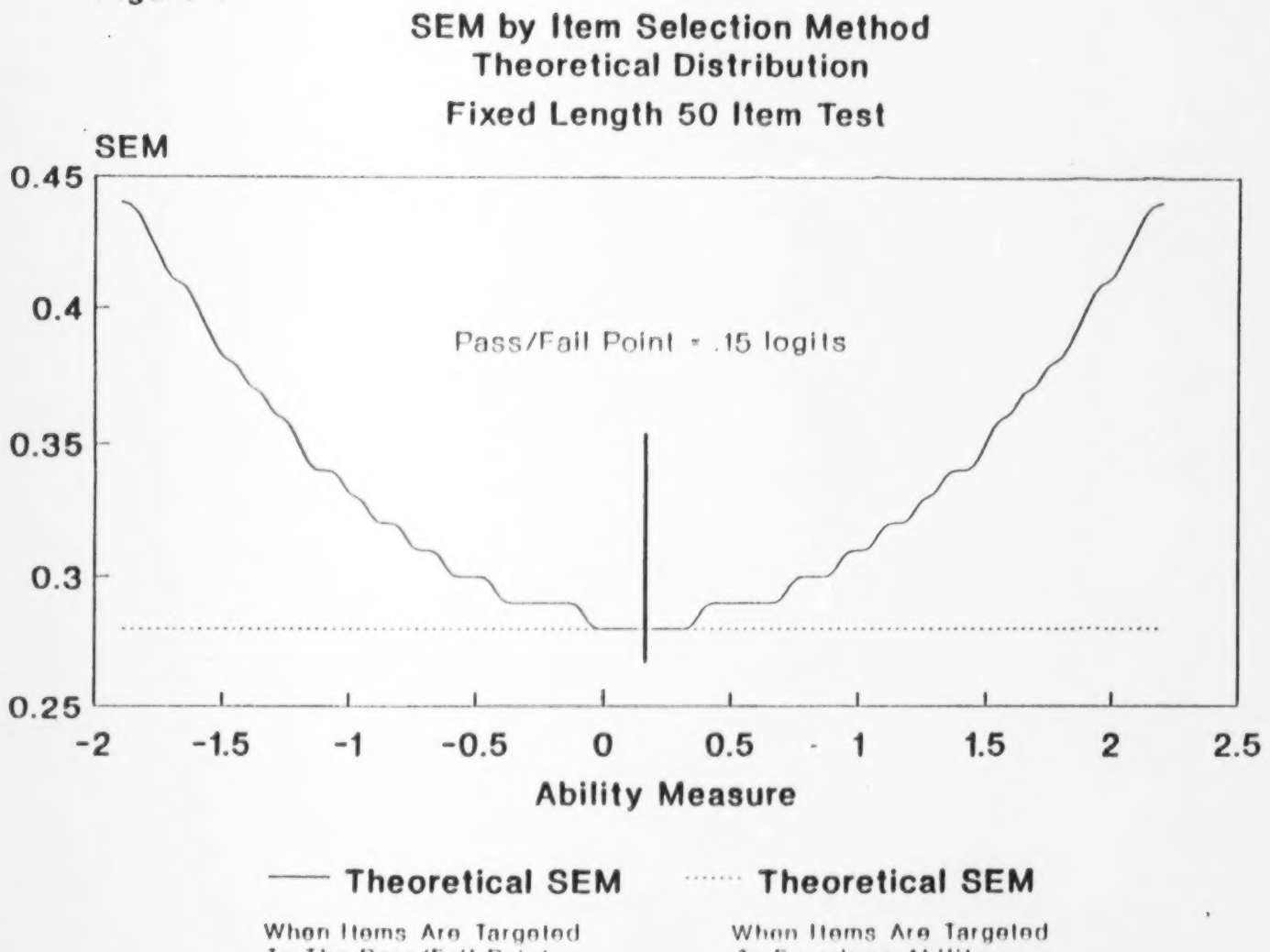
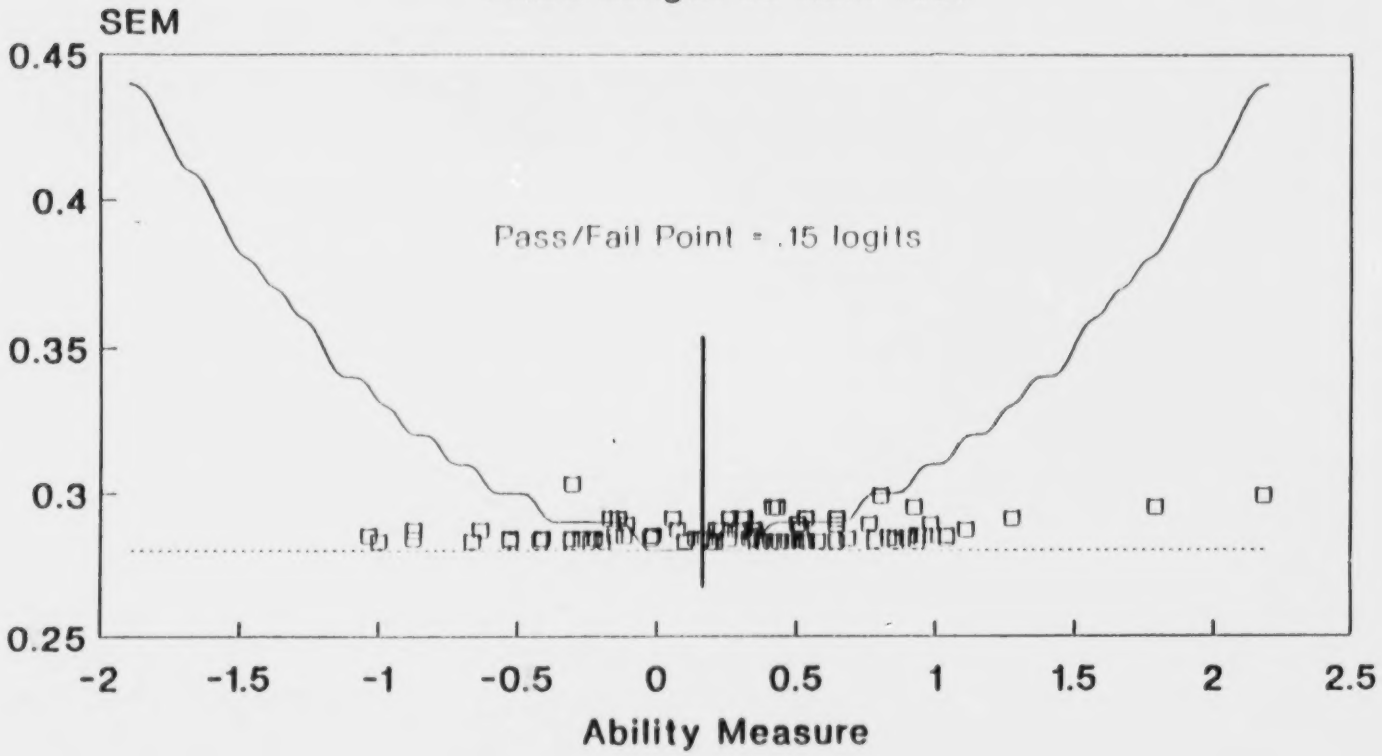

SEM by Item Selection Method
Theoretical Distribution
Fixed Length 50 Item Test
SEM
0.45
0.4
0.35
0.3
0.25
Pass/Fail Point = .15 logits
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Ability Measure
Theoretical SEM
When Items Are Targeted
To The Pass/Fail Point
Theoretical SEM
When Items Are Targeted
To Examinee Ability

Figure 2

SEM by Item Selection Method
Theoretical and Observed Distributions
Fixed Length 50 Item Test



Theoretical SEM When Items Are Targeted To The Pass/Fail Point

Theoretical SEM When Items Are Targeted To Examinee Ability

Observed SEM When Items Are Targeted to Estimated Examinee Ability

Figure 3

SEM by Item Selection Method

Theoretical and Observed Distributions

Variable Length/Maximum Test 100 Items